Phase I Project Summary

Firm: Intelligent Automation, Inc.
Contract Number: NNX13CA44P
Project Title: Cloud-based Analytics, Store and Query System (CASQUE) for Data-intensive Scientific Processing

Identification and Significance of Innovation: (Limit 200 words or 2,000 characters whichever is less)

Many NASA missions, e.g., Kepler or Operation IceBridge, involve collecting massive amount of data, which might amount to petabytes during the lifetime of the mission. Typically, the data in its raw form is shipped to a processing center, and after various levels of processing, it is made available to scientists for analysis and prediction. However, is time-consuming, and neither cost-effective nor scalable (elastic). Cloud computing offers a promising environment for virtualized computation and storage (complementary to High Performance Computing) at different levels to analyze the massive amounts of data created by NASA missions: as an infrastructure (e.g., Amazon Web Services (AWS)), as a platform (e.g., Google Apps), and as a software ecosystem (e.g., MapReduce, its implementation Hadoop and NoSQL tools). NASA, as part of its Big Data vision, e.g., in Open Government 2.0, can leverage similar cloud capabilities for the benefit of the research and science community via open-source data and software, and CASQUE is proposed as a data-intensive scientific computing framework that leverages cloud and is specialized for NASA missions.

Technical Objectives and Work Plan: (Limit 200 words or 2,000 characters whichever is less)

- **Objective 1: Develop the cloud-based workflow and architecture for a use case.** One major effort in the Phase I project was to develop the workflow and the architecture for a use. This effort was based on inputs from NASA scientists and mission dependent. We also studied the trade-offs of various workflows in terms of how data is managed, archived and queried, and incorporate into our design.
- **Objective 2: Implement the prototype tailored for a NASA mission.** This objective was based on the chosen mission, the datasets, and related analytics. Based on our discussions, we chose parallelization of "unsupervised outlier detection" as the mission of interest. We used both MapReduce and MPI-based programming, and integrated them to parallelize the algorithm.
- **Objective 3: Large-scale testing.** We tested the system in IAI premises with IAI's private cloud and configured VMs. The performance metrics such as speed-up time, threshold, and pruned number of points were evaluated at this stage.

Technical Accomplishments: (Limit 200 words or 2,000 characters whichever is less)

In our Phase I efforts, we focused on demonstrating the feasibility of a cloud-based Big Data analytics, store and query for data-intensive scientific computing using "distributed implementation of outlier detection iOrca" as the major use case. First, we studied the feasibility of using cloud-based analytics for select NASA missions. Second, we designed and implemented unsupervised parallel outlier detection. Third, we tested the performance of the unsupervised outlier detection algorithm using datasets up to 1e5 points with 7 to 10 dimensions. Finally, based on the outcomes of the Phase I, together with the NASA ARC team, we have decided what

further development unsupervised outlier detection data product would require, and what other applications (and datasets) would benefit from this development.

NASA Application(s): (Limit 100 words or 1,000 characters whichever is less)

NASA, as part of its Big Data vision, e.g., in Open Government 2.0, can leverage similar cloud capabilities for the benefit of the research and science community via open-source data and software. There are many data mining applications currently hosted in DashLink and NEX that can be converted into cloud-based analytics. Also, typical airborne missions, e.g., Operation IceBridge and EcoSAR, will benefit tremendously from a seamless cloud-based store and analyze system. Similarly, the seamless submission and incorporation of diverse data, open access data and information exchange, and open development of open source software and tools will allow foster the data-intensive research dependent on the success of NASA missions.

Non-NASA Commercial Application(s): (Limit 200 words or 2,000 characters whichever is less)

Cloud-based analytics have immediate applications in DoD, especially in the intelligence community. As part of the Naval Tactical Cloud,DCGS-N and DCGS-MCare two efforts, where we could provide increased (SA) by directly using the cloud-based analytics. In banking and finance, the developed outlier detection algorithms can speed-up detecting fraud and insider trading.

Name and Address of Principal Investigator: (Name, Organization, Street, City, State, Zip)

Onur Savas, Ph.D.
15400 Calhoun Drive, Suite 400, Rockville, MD 20855

Name and Address of Offeror: (Firm, Street, City, State, Zip)

Intelligent Automation, Inc.
15400 Calhoun Drive, Suite 400, Rockville, MD 20855